Virus Research xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

On the biased nucleotide composition of the human coronavirus RNA genome

Ben Berkhout*, Formijn van Hemert

Laboratory of Experimental Virology, Department of Medical Microbiology, Center for Infection and Immunity Amsterdam (CINIMA), Academic Medical Center, University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Coronavirus
RNA genome
Nucleotide signature
MERS
SARS

ABSTRACT

We investigated the nucleotide composition of the RNA genome of the six human coronaviruses. Some general coronavirus characteristics were apparent (e.g. high U, low C count), but we also detected species-specific signatures. Most strikingly, the high U and low C proportions are quite variable and act like communicating vessels, C goes down when U goes up and vice versa. U ranges among virus isolates from 30.7% to 40.3%, and C makes the opposite movement from 20.0% to 12.9%, respectively. The nucleotide biases are more pronounced in the unpaired regions of the structured RNA genome, which may suggest a certain biological function for these distinctive sequence signatures. Coronaviruses have an atypical codon usage that has been linked to mutational events operating on the viral RNA genome on an evolutionary time scale. We suggest that the atypical nucleotide bias may serve a distinct biological function and that it is the direct cause of the characteristic codon usage in these viruses. The relevance for evolution of the novel human pathogens MERS and SARS is discussed.

1. Introduction

Coronaviruses are positive sense, single-stranded RNA viruses that infect a wide range of animals. The coronaviridae can cause a spectrum of diseases, ranging from respiratory, enteric, hepatic and neurological diseases of varying severity. Two human coronaviruses that cause relatively mild respiratory symptoms are known since the 1960s: HCoV-229E and HCoV-OC43. SARS-CoV was identified in 2003 and causes a more severe respiratory syndrome (Fouchier et al., 2003). The fourth member of the coronaviridae family that was identified in 2004 to cause respiratory symptoms in humans is HCoV-NL63 (van der Hoek et al., 2004). The fifth member HCoV-HKU was described the next year (Woo et al., 2005). More recently, the pathogenic MERS coronavirus was identified in the Middle East as the sixth human coronavirus (Zaki et al., 2012). Both SARS and MERS represent recent zoonotic transfers into the human population. The full-length sequences of the RNA genomes of all these human coronaviruses have been analyzed to some extent (van Boheemen et al., 2012; Woo et al., 2007; Pyrc et al., 2004, 2007; Marra et al., 2003). The coronavirus RNA genome is around 30 kb and thereby the largest among the profoundly diverse group of RNA viruses. In global terms, a similar genome organization is apparent for all coronaviruses and a similar set of proteins is encoded. Besides encoding for viral proteins, the viral RNA genome usually contains multiple molecular signals, either RNA sequence elements or secondary and higher order RNA structures that interact with viral or cellular components – proteins or RNAs – to facilitate certain steps of the viral replication cycle. This is also true for the coronaviridae, which for instance encode the Transcription Regulation Sequence (TRS) that is involved in the induction of a discontinuous transcription mechanism to generate subgenomic mRNAs that encode the different viral proteins. In this study, we want to focus on a more basic property of the coronavirus RNA genomes: their biased nucleotide composition.

There have been previous reports on the biased nucleotide composition of coronavirus RNA genomes. Grigoriev performed a cumulative skew analysis to analyze mutational patterns and reported an excess of G compared with C, suggestive of excessive C-to-U deamination among the coronaviruses, but significantly less for SARS (Grigoriev, 2004). A subsequent analysis based on the NL63 genome indicated a very low C count and a high U count (Pyrc et al., 2004). Both studies reported a clear difference in the magnitude of the nucleotide bias between the first two-third and last one-third of the coronavirus genome, which likely relates to the mechanism of subgenomic mRNA synthesis and exposure of single-stranded

* Corresponding author. Tel.: +31 205664822.
E-mail address: b.berkhout@amc.uva.nl (B. Berkhout).

http://dx.doi.org/10.1016/j.virusres.2014.11.031
0168-1702/© 2015 Elsevier B.V. All rights reserved.

RNA domains. Cytosine deamination and discrimination against CpG dinucleotides were proposed as the driving forces that shaped the coronavirus RNA genomes over evolutionary times (Woo et al., 2007, 2010). Related to this deviant nucleotide count, the codon usage of the coronaviruses is also particularly unusual (Gu et al., 2004).

Several new findings urged us to revisit this topic. First, the MERS coronavirus as novel human pathogen follows some general coronavirus trends, but is also characterized by some unique and rather extreme features of nucleotide usage. Second, we can present a simple, but striking classification of the human coronaviruses based on their nucleotide composition. Coronaviral RNA genomes have a rather stable G and A count, but vary significantly in the U and C distribution, with as two extremes MERS (32.5% U, 20.3% C) and HKU (40.3% U, 12.9% C). Third, we performed for the first time a nucleotide usage analysis in the context of the structured coronavirus RNA genomes. All this information provides new mechanistic insight. Specifically, some of the previously proposed mutational scenario's (e.g. CpG discrimination) become less likely and we propose alternatives for the C deamination scenario that involves a cellular cytosine deaminase enzyme. Specifically, the differential U versus C bias among the human coronaviridae may suggest a mutational property of the diverse viral polymerases. Alternatively, the specific nucleotide composition of the viral RNA genomes may have been selected to execute a certain biological function. Finally, this new insight also strongly influences the way we should look at the atypical codons used by these viruses.

2. Materials and methods

Nucleotide sequences of the coronaviruses were taken from Genbank (MERS: JX869059, SARS: NC004718, CoV 229E: KF514433, CoV OC43: NC005147, CoV NL63: JX504050, HKU-1A: DQ415914, HKU-1B: DQ415911, HKU-1C: DQ415912). MFold (Zuker and Turner, 1999) was used with default settings for RNA secondary structure prediction. The single-stranded or ss-count file of an MFold output supplied the number of folded structures (50 maximally), including a frequency value for each individual nucleotide of being unpaired in this collection of structures. We scored a nucleotide as unpaired (single-stranded, "ss") if half or more than half of the structure models reported its position as "ss". Nucleotides with a ss-count value below this criterion were scored as being paired (double-stranded, "ds"). Discrimination between ss and ds nucleotides was performed in Excel and fasta files were created in order to determine the composition of ss and ds nucleotides, separately, by means of MEGAv5 (Tamura et al., 2011).

The size limit for submission to the MFold server is 9000 nucleotides (nts). The 30,000 nts coronaviral RNA genomes were therefore partitioned into four portions (3 × 8500 nts and rest) with 500 nts overlaps. This obviously ignores long-distance interactions that may occur in a coronavirus RNA genome. The ss-count data of the submission output files were arithmetically averaged at the region of overlap before the ss/ds discrimination was performed. We used the MFold ct file of the top 1 structure model to score the basepair usage in coronaviral RNA (regular Watson–Crick and G-U/U-G pairs). Partial sequence files were reconstituted at a site near the center of the overlap to minimize folding artifacts near the borders of the submitted sequences.

Base composition analysis along the RNA genome length and the accompanying ss and ds fasta files was performed by the method of cumulative skew diagrams in overlapping windows (Grigoriev, 1998). For normalization purposes, windows were defined around 1% of the sequence length with a step size of 20% of the window size.

Codon usage was characterized by means of plotting the effective number of codons (ENC-values) of coronavirus genes versus their GC-content at the 3rd synonymous codon positions (GC3-values): the "Nc-plot" (Wright, 1990). This analysis excludes the codons AUG (Met) and UGG (Trp). A continuous line indicates theoretical ENC values with random codon usage as a function of GC3. Deviation from this line in the direction of lower ENC-values points to translational selection acting in favor of a preferred set of codons, as has been described for highly expressed genes in yeast (Bennetzen and Hall, 1982) and Escherichia coli (Sharp and Li, 1987). Codon usage data for the nuclear genes of the host species were obtained from the Codon Usage Database (Nakamura et al., 2000). The data were available for the following numbers of codons (number of available coding sequences or CDS in parentheses): human: 40,662,582 (93,487), dromedary: 6414 (21) and bats (3 species): 3522 (10).

Calculations were performed in Excel.

3. Results

3.1. The nucleotide count of the RNA genome differs per coronavirus

Table 1 lists the nucleotide count of the human coronavirus RNA genomes, ranked from highest C-count (MERS, 20.3%) to the lowest (HKU, 12.9%). We included 3 HKU isolates (1A, 1B, 1C) to demonstrate conservation of these nucleotide characteristics among virus isolates of a particular coronavirus. Strain-specific trends are also conserved for different isolates of the other coronaviruses (not shown). MERS and SARS, which represent two recent zoonotic transmissions into the human population, are both present on the same end of the spectrum. SARS is quite extreme with a C-count of 20.0% and in fact the lowest U-count of 30.7%. The highest count of 40.3% U is apparent for the 1B isolate of HKU. These numbers seem quite dramatic. For instance, the extremely biased HIV-1 RNA genome reaches a maximal A-count of 36.7% (van der Kuyl and Berkhout, 2012).

Some intriguing patterns become apparent by inspection of the nucleotide counts of the human coronaviruses. A few general coronavirus rules are observed. The first relates to the pyrimidines: the U-count is above average and the C-count is below average. The second rule applies to the purines and is less prominent, but A is preferred over G. There are also species-specific trends. In particular, the C/U ratio differs profoundly per coronavirus type, and these nucleotides seem to behave as communicating vessels. To illustrate that C and U seem to be competing for sequence space, we plotted the nucleotide composition per coronavirus (color coded) in Fig. 1. This picture also nicely illustrates that most variation occurs in the C/U and not the A/G section. The A/G ratio is rather stable among different coronaviruses, with minor fluctuations in the A-count (ranging from 26.2% for MERS to 28.5% for SARS) and the G-count (ranging from 19.0% for HKU-1A to 21.7% for OC43). For the more detailed follow-up analyses, we selected the two extremes MERS (32.5% U) and HKU (the 1B isolate with 40.3% U, simply called HKU hereafter).

Table 1
Differential nucleotide composition among coronaviruses.

Coronavirus	ID	A	U	C	G
MERS	JX869059	26.2	32.5	20.3	20.9
SARS	NC_004718	28.5	30.7	20.0	20.8
229E	KF514433	27.1	34.7	16.6	21.6
OC43	NC_005147	27.6	35.6	15.2	21.7
NL63	JX504050	26.3	39.2	14.4	20.1
HKU-1C	DQ415912	27.8	40.1	13.0	19.1
HKU-1A	DQ415914	27.9	40.2	13.0	19.0
HKU-1B	DQ415911	27.7	40.3	12.9	19.1

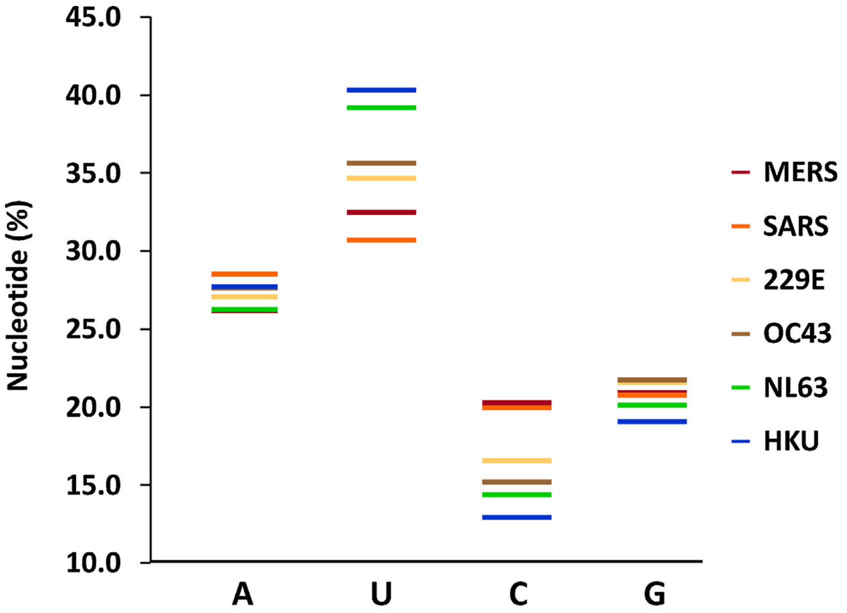

Fig. 1. Summary of the nucleotide composition of coronavirus RNA genomes. A and G proportions are relatively invariant among coronaviruses, while in contrast U and C are highly variable and represent communicating vessels.

3.2. Nucleotide distribution in structured coronavirus RNA

For HIV-1 RNA, we recently analyzed the nucleotide composition in the context of the structured RNA genome because an experimentally probed RNA secondary structure model was available (van Hemert et al., 2013; Watts et al., 2009). In particular, we were interested to map the distribution of the different nucleotides in the single-stranded and double-stranded parts across the genome. We described that the HIV-specific nucleotide bias is even more extreme in the unpaired regions of HIV-1 RNA (van Hemert et al., 2013) and subsequently demonstrated that the same trend is apparent for MFold-predicted RNA structures (Van Hemert et al., 2014). As no experimentally probed RNA structure models are available for the complete coronavirus genomes, we relied on computer-generated RNA structures in a first attempt to analyze the structural presentation of the different nucleotides.

The RNA genomes of the two extremes, MERS (relatively low U) and HKU (relatively high U), were folded with the MFold program using default settings to yield 30–50 structures. We subsequently investigated the predicted structures of the RNA genomes of other coronaviruses. Table 2 lists the nucleotide composition of the unpaired (single-stranded or ss) and basepaired (double-stranded or ds) nucleotides. The trend first described for HIV-1 RNA that the extremes become more extreme in the ss domains and consequently less extreme in the ds domains is also apparent for the HKU RNA genome. The high U-count of HKU (40.3%) goes up to 48.3% for ss and down to 36.1% for ds nucleotides. These values are extreme, but HIV-1 reaches up to 50.3% A in ss regions of its RNA genome. The C-count makes the contrasting movement: from a mere 12.9% further down to 11.0% in ss and up to 13.9% in ds domains. Thus, the nucleotide composition is put to the extreme for ss regions and approaches more neutral values in the ds regions.

The situation is more complex for MERS, where the relatively suppressed U-count (only 32.5% compared to 40.3% for HKU) is neither suppressed more severely in the ss domains (38.6%) nor overrepresented in the ds domains (29.1%). The C-count is relatively stable in both compartments. The general value of 20.3% becomes 19.4% for ss and 20.8% for ds nucleotides.

3.3. The nucleotide composition influences the basepair usage in coronavirus RNA

We next probed whether the atypical nucleotide composition does influence the type of basepairs used in the structured RNA genomes. The previous HIV-1 study noticed a trend toward the more frequent usage of the more stable basepairs (GC and CG > AU and UA > GU and UG), which correlated with the G and C preference in the ds parts of the HIV-1 RNA genome (van Hemert et al., 2013). We now analyzed the coronavirus RNA genomes. As said, we focused on the most extreme MERS and HKU genomes and we used the A-rich HIV-1 RNA as outgroup for comparison.

The major conclusion is that the basepair composition correlates with the biased nucleotide counts (Fig. 2). The general coronavirus pattern (U up, C down) is visible in the basepairs: G–U/U–G and U–A/A–U are relatively up and G–C/C–G are down, which is especially true for the most extreme HKU genome. The HKU genome is

Table 2
Nucleotide composition in RNA structure models of MERS and HKU genomes.

Coronavirus	ID		A	U	C	G	nts
HKU-1B	DQ415911	All nts	27.7	40.3	12.9	19.1	29,904
		ss nts	28.1	48.3	11.0	12.5	10,154
		ds nts	27.5	36.1	13.9	22.5	19,748
MERS	JX869059	All nts	26.2	32.5	20.3	20.9	30,119
		ss nts	31.2	38.6	19.4	10.8	10,948
		ds nts	23.3	29.1	20.8	26.7	19,173

Coronavirus RNA (30,000 nts) was divided into 4 parts (3 × 8500 + rest) with a 500 nts overlap allowing reconstitution before analysis. The ss-count output file of Mfold (based on max 50 structures) was used for data calculation.

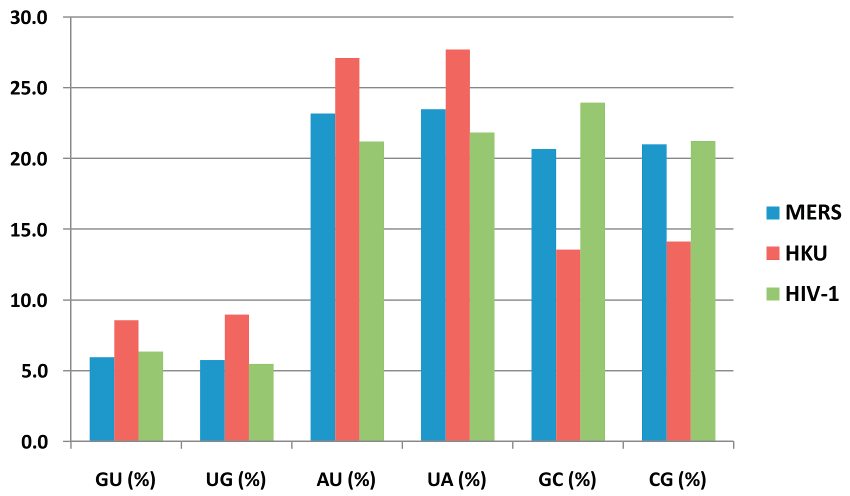

Fig. 2. Basepair composition in RNA structure model of MERS, HKU and HIV-1. Only the top MFold predictions were analyzed. Coronavirus RNA genomes (30,000 nts) were split in four fragments (3 × 8500 + rest) with 500-nts overlaps.

low in ds C (13.9%), but remains fairly high in ds G (22.5%), which means that more G–U/U–G basepairs must have been accomodated. U still prefers to pair with A (53.9%), but we scored a significant number of pairings with G (19.5%). Overall, a rather distinct basepair composition is apparent for HKU coronavirus compared to HIV-1. The biggest difference is apparent for G–C pairs: 26.5% for HKU and 45.1% for HIV-1. But we also observed notable differences between MERS and HKU that relate to the different nucleotide composition of their RNA genomes. The basepair composition of MERS coronavirus resembles that of HIV-1 RNA much more than that of HKU.

3.4. Skew analysis along the coronavirus RNA genome

We performed a nucleotide skew analysis along the MERS and HKU genomes to reveal the fine structure of the ss and ds segments (Fig. 3). An advantage of a skew analysis is that it allows one to score global trends across genomes by minimizing local fluctuations. Skew values were plotted for all six nucleotide comparisons (G vs A, C vs A, U vs A, U vs C, C vs G, U vs G) along the 30 kb viral genome in overlapping windows of 1% of the sequence size with a step size of 20% of the window size. It should be noted that GA in skew language does not represent a basepair, but a comparison of the number of G-nucleotides with the number of A-nucleotides. As a result, about 500 data points were obtained comprising each *X*-axis irrespective of the length of the nucleotide sequence involved. We specifically used the same scale on the *Y*-axis to allow a direct comparison of the virus-specific signatures.

The general skew analysis in Fig. 3 (left panels) is in line with the general coronavirus pattern (U up, C down), with trends that pertain across the genome. U wins from the other three nucleotides as evidenced by the ascending lines, most pronounced from C, than G and A. This holds true even for the more moderately biased MERS RNA. As expected, the HKU skew is much more extreme (steeper lines) than that of MERS, but the patterns are similar. The skew patterns confirm that there is relatively much variation in the U/C values and little variation in the G/A ratio.

We next performed a skew analysis for the ds and ss nucleotides (middle and right panels, respectively). The results confirm that the bias is more extreme (steeper skew lines) for the unpaired (ss) nucleotides. Skew values are relatively large in the ss domains, relatively small in the ds domains and, as expected, intermediate

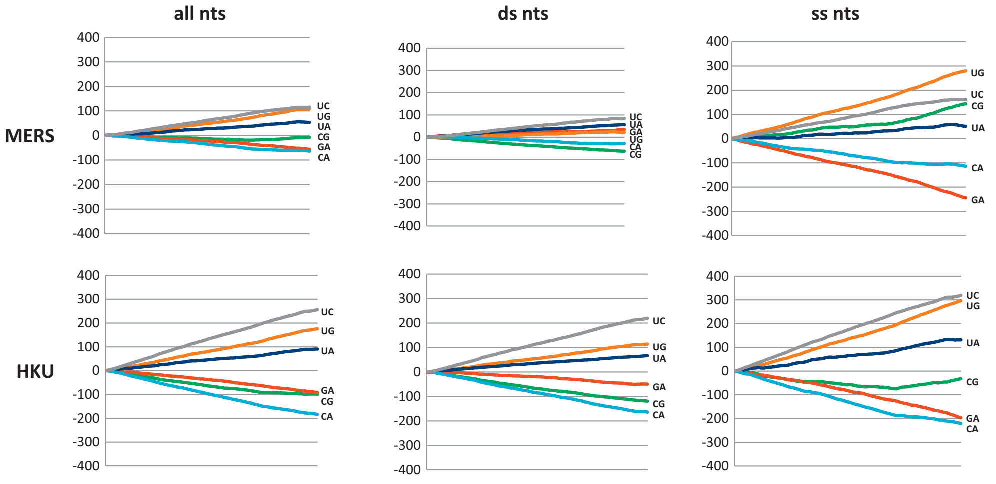

Fig. 3. Skew analysis of RNA genomes of MERS and HKU. Skew values $(N1-N2)/(N1+N2)$ have been calculated in overlapping windows along the sequence ("all nts", "ds nts" and "ss nts"). Window size was set at 1% of the length of the sequence with a step size of 20% of the window size resulting in approximately 500 datapoints comprising the *X*-axis. We used the same *Y*-axis for the cumulative skew values to allow a direct comparison of the compositional signatures of different coronavirus RNA genomes. It should be noted that the labels next to each line do not indicate a basepair, but refer to the two nucleotides for which the skew values were calculated.

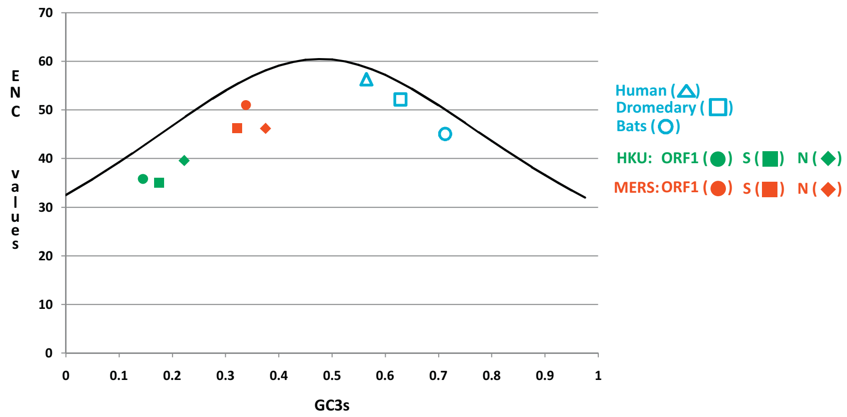

Fig. 4. Codon ENC analysis of MERS and HKU. The effective number of codons (ENC-values, *Y*-axis) of coronavirus genes was plotted against the GC-content at the 3rd synonymous codon positions (GC3-values, *X*-axis). The continuous line indicates theoretical ENC values with random codon usage as a function of GC3. Deviation from this line in the direction of lower ENC-values points to translational selection acting in favor of a preferred set of codons. Codon usage data for the nuclear genes of the host species were obtained from the following numbers of codons (number of genes in parentheses): human: 40,662,582 (93,487), dromedary: 6414 (21) and bats (3 species): 3522 (10).

for all nucleotides. All skews are in agreement with the individual nucleotide composition of these viruses.

The coronavirus skew values, as soon as they differ significantly from zero, are represented by straight lines along the genome length, indicating that they represent an intrinsic property of the viral genome that is not restricted to specific parts of the viral genome. In other words, the observed biases represent a stable property, but one noticeable exception seems to be present. A shift is apparent at two-third of the MERS and HKU genome length. This shift occurs in a region between the 1A/1B and S genes on the viral RNA genome. The shift is visible in all nucleotide analyses of MERS, but appears more extreme in the ss skews. A more dramatic switch of the direction of the CG skew line is apparent for HKU. Two possible mechanisms have been proposed in literature to explain this shift or switch in the nucleotide bias (Grigoriev, 2004; Pyrc et al., 2004).

3.5. Codon usage is dictated by the nucleotide composition of coronaviruses

We previously described that the A-rich HIV-1 RNA dictates the exotic codon usage of this virus. We therefore wanted to know whether the particular codon usage of coronaviruses, as described in literature, correlates with the nucleotide biases. We performed an analysis of the effective number of codons (ENC) used by the two extreme genomes of MERS and HKU. We compared the G and C content of the synonymous third codon positions (GC3) in the coronavirus gene 1 (ORF1, occupying the first two-third of the viral genome) vs the S and N genes in the last one-third, thus downstream of the observed shift/switch in the skew analysis (Fig. 4). The 3′-located S and N genes do not behave significantly different from the 5′-located gene 1. The three genes cluster together for each coronavirus species, but HKU is more extreme than MERS toward low GC3 values, as expected. The codons do not deviate much from the bell-shaped line that represents ENC values of random codon usage expected for a given GC3 composition. The characteristic distinction between MERS and HKU is the difference of GC content at the 3rd synonymous codon position. Thus, the codons follow the nucleotide count and the overall codon usage is not biased in other ways.

A gross difference is apparent for codon usage in a few possible hosts: human, bat and dromedary as candidate host for MERS (Haagmans et al., 2014). We performed additional analyses with all coronaviruses listed in Table 1. GA is stable among the different coronaviruses and UG was used to visualize U-preference (results not shown), but none of the coronaviruses cluster with any of the hosts. There is variation among coronaviruses due to the magnitude of the U/C-bias: MERS is least extreme and HKU most extreme. All other human coronaviruses occupy in between spots in the CG3 plot (not shown), consistent with their intermediate nucleotide count. These relationships hold perfectly for the longest ORF1. A few exceptions were apparent for the much smaller and thus less reliable ORFs encoding the S protein (229E a bit more extreme than MERS) and the N protein (SARS and OC43 more extreme than MERS). Most importantly, all coronavirus genes are positioned rather close to the bell-shaped curve, which indicates the absence of selection of certain “preferred” codons, and relatively far away from the GC3 values of any of the candidate host species. We conclude that the particular codon usage is determined largely by the biased nucleotide count of the coronavirus genomes.

3.6. A more detailed codon analysis

We also inspected the detailed codon tables that have been presented by others (Gu et al., 2004; Woo et al., 2007) for trends that could support our analyses. We generated a detailed survey of the codons used in the largest ORF1 in HKU vs MERS (Supplementary Table S1) and calculated the nt-count per codon position (Supplementary Table S2). Indeed, the first general coronavirus rule about pyrimidines (U up, C down) dictates codon usage in all codon groups and at the three codon positions without any exceptions. This rule holds for the broad collection of human and animal coronaviruses that were analyzed by Woo et al. (2007), and among the human coronaviruses the effect ranges from MERS (most modest) to HKU (most extreme). This rule is perhaps most dramatically visualized for the 2-codon groups where the choice is C or U. Phenylalanine is encoded by UUU or UUC, but a 0.94/0.06 = 15.7-fold bias for U is present in HKU, contrasting with a modest 1.7-fold bias in MERS. All other coronaviruses take an intermediate position. Another example is provided by the 4-codon groups. Valine is encoded by the four codons GUN, which are used with profound different frequencies in HKU: 0.68 GUU, 0.05 GUC, 0.20 GUA and 0.06 GUG, thus yielding a 13.6-fold bias for U over C, but a much more modest 2.2-fold bias in MERS. Also, the coronaviral rule (U up, C down) is more prominent in HKU than in MERS and hence, different proportions of ORF1 encoded amino acids with the C-nucleotide at the 2nd codon position can be expected (Supplementary Table S2). Indeed, HKU/MERS ratios of 0.86 (Ser), 0.88 (Pro), 0.79 (Thr) and 0.76 (Ala)

Please cite this article in press as: Berkhout, B., van Hemert, F., On the biased nucleotide composition of the human coronavirus RNA genome. Virus Res. (2015), http://dx.doi.org/10.1016/j.virusres.2014.11.031

are apparent (Supplementary Table S1). Nucleotide preferences of coronaviruses (U up, C down) affect the amino acid composition of the viral proteins.

Supplementary Table S1 related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.virusres.2014.11.031.

Supplementary Table S2 related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.virusres.2014.11.031.

The second general coronavirus rule about purines (A over G) is also well supported, but with less dramatic numbers. For A/G choices, A always wins in the coronaviruses. For instance, lysine is encoded by AAA or AAG that are used in quite different fraction of codons (0.68 vs 0.32) in HKU, yielding a 2.1-fold bias of A over G. More neutral values are apparent for MERS (1.1-fold bias of G over A). Very similar values are observed for A/G choices. All these findings follow the basic nucleotide count in these genomes as illustrated in Fig. 1.

4. Discussion

We analyzed the nucleotide composition of the RNA genome of human coronaviruses and arrive at some general and species-specific rules. Two general coronavirus rules are apparent that relate to the usage of pyrimidines (C over U) and purines (A over G). The A/G bias is a relatively stable property among the coronaviruses. In contrast, the C/U bias differs significantly per virus type and we scored U-counts from 30.7% (SARS) to 40.3% (HKU) and C-counts from 20.3% (MERS) to 12.9% (HKU). The C- and U-counts behave as communicating vessels. Although this study was restricted to the human coronaviruses, these basic properties apply to all known animal and human coronas (results not shown). A quick survey revealed a new record number for the Bat-SARS-CoV with 30.5% U (Woo et al., 2007). Perhaps surprisingly, we think that these basic nucleotide trends have not been reported previously. Although they may seem useless numbers for some, we think that these basic properties can encode important biological functions and one may start wondering about the evolutionary history of these sometimes striking nucleotide features. The biased nucleotide composition can also have a major influence on derived parameters. For instance, our analysis clearly suggests that the nucleotide composition largely dictates the codons that are used by these viruses for the translation of the RNA genome and sub-genomic mRNAs.

Previous studies on codon usage in different viruses have highlighted mutational pressure as the major factor in shaping codon usage patterns compared with natural selection (Gu et al., 2004; Jenkins and Holmes, 2003a; Wang et al., 2011; Wong et al., 2010; Woo et al., 2007; Gu et al., 2004). For coronaviruses, cytosine deamination and the selection against CpG motifs have been proposed as the mutational forces that shaped the viral genome and its codons (Grigoriev, 2004; Woo et al., 2007; Pyrc et al., 2004). However, as our understanding of codon usage increases, it appears that although mutational pressure is still a major driving force, it is certainly not the only force when considering different types of RNA and DNA viruses (Berkhout and van Hemert, 1994; van Hemert and Berkhout, 1995; Chen, 2013; Shi et al., 2013; Zhang et al., 2013). For HIV-1 with its A-rich RNA genome, we initially proposed two evolutionary scenarios. Mutations could be introduced on an evolutionary time scale by the error-prone reverse transcriptase or cellular enzymes like Apobec, but we also stressed that these atypical RNA molecules may have been selected to exert a certain biological function. For instance, the distinctive nucleotide composition influences the overall folding of the RNA molecule and may thus affect specific replication steps like packaging of the RNA genome in virion particles. The genome composition may also relate to the intensive virus-host interaction, e.g. by avoiding recognition by the innate immune system. For HIV-1 with its extremely A-rich genome it was recently proposed that this property helps to avoid recognition by the innate immune system (Vabret et al., 2012), which could provide strong selective pressure on retroviruses and many RNA viruses including coronaviruses (van der Kuyl and Berkhout, 2012; Van Hemert et al., 2014; Kindler and Thiel, 2014). There could also be a more passive function for the biased genome composition. For HIV-1, A-rich sequences may restrict the number of sites that can be mutated and inactivated by the cellular Apobec restriction factor. For coronaviruses, the C-rich genome may highlight some important replication signals such as the A-rich TRS element (e.g. AACUAAA in NL63 (Pyrc et al., 2004)) that is positioned at several locations within the coronaviral genome.

Our finding that these nucleotide signatures are even more pronounced in the single-stranded domains of the coronaviral genomes perhaps support this latter selection theory. The nucleotide trends are certainly not restricted to the coding regions of these genomes as they are also apparent in the non-coding 5′UTR, indicating that these signatures are not invented to create a certain codon bias and that the effect is executed at the level of translation, but rather that it serves another biological purpose. In fact, this nucleotide bias also directly influences many other parameters such as the dinucleotide composition and possibly even the amino acid composition of the encoded viral proteins (Berkhout and van Hemert, 1994). We previously indicated that serious nucleotide skews may even trigger phylogenetic artifacts: viruses with a similar nucleotide preference tend to cluster, but are not necessarily related by descent (van Hemert and Berkhout, 1995). We also cannot formally exclude that the bias operates at the level of the genomic minus-strand RNA, which obviously has the opposite characteristics (C over U becomes G over A).

Do these results tell us something about virus pathogenicity and evolutionary events during zoonotic transmissions? Although the two serious pathogens SARS and MERS are present on one side of the C/U spectrum with a relatively low U-count and high C-count, it seems dangerous to propose a correlation between the nucleotide signature and pathogenicity as this may just be a coincidence. Most viral nucleotide/codon characteristics appear not to depend on the host organism as we observed similar properties for murine, avian, bat and human coronaviruses (results not shown). The codon analysis presented in Fig. 4 also clearly indicates that there is no virus adaptation to the host, at least in this respect. One evolutionary scenario that links nucleotide usage to pathogenicity remains possible. The new coronaviruses that arrived in humans via zoonotic transfer are pathogenic (MERS, SARS). The ones that are circulating in humans for a much longer period may have adapted to become less pathogenic. This idea is similar to a natural attenuation scenario that has been proposed for other viruses like myxoma virus in a new Australian epidemic among the introduced European rabbits (Kerr et al., 2012). It is beneficial for the virus not to kill the host too quickly as this increases the chance of viral spread. We may see this adaptation as a gradual increase of U and decrease of C, which may attenuate viral gene expression (e.g. sub-optimal codon usage, visible as gradual deviation from the ENC curve, or by another mechanism: e.g. reduced RNA packaging capacity or increased recognition by the innate immune system). This hypothesis predicts that MERS and SARS will evolve to become more U-rich and C-poor in humans, but that obviously requires the viral presence in the human population over evolutionary times.

Finally, knowledge of the nucleotide and codon usage in viruses can not only reveal information about molecular evolution, but also improve our understanding of the regulation of viral gene expression and aid vaccine design, e.g. by providing novel ways for stable virus attenuation.

Acknowledgement

This research is sponsored by NWO (700.59.301) (TOP grant to BB).

References

Bennetzen, J.L., Hall, B.D., 1982. Codon selection in yeast. J. Biol. Chem. 257, 3026–3031.

Berkhout, B., van Hemert, F.J., 1994. The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. Nucl. Acids Res. 22, 1705–1711.

Chen, Y., 2013. A comparison of synonymous codon usage bias patterns in DNA and RNA virus genomes: quantifying the relative importance of mutational pressure and natural selection. Biomed. Res. Int. 2013, 406342.

Fouchier, R.A., Kuiken, T., Schutten, M., Van Amerongen, G., van Doornum, G.J., van den Hoogen, B.G., Peiris, M., Lim, W., Stohr, K., Osterhaus, A.D., 2003. Aetiology: Koch's postulates fulfilled for SARS virus. Nature 423, 240.

Grigoriev, A., 1998. Analyzing genomes with cumulative skew diagrams. Nucl. Acids Res. 26, 2286–2290.

Grigoriev, A., 2004. Mutational patterns correlate with genome organization in SARS and other coronaviruses. Trends Genet. 20, 131–135.

Gu, W., Zhou, T., Ma, J., Sun, X., Lu, Z., 2004. Analysis of synonymous codon usage in SARS coronavirus and other viruses in the Nidovirales. Virus Res. 101, 155–161.

Haagmans, B.L., et al., 2014. Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. Lancet Infect. Dis. 14, 140–145.

Jenkins, G.M., Holmes, E.C., 2003a. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 92, 1–7.

Kerr, P.J., Ghedin, E., DePasse, J.V., Fitch, A., Cattadori, I.M., Hudson, P.J., Tscharke, D.C., Read, A.F., Holmes, E.C., 2012. Evolutionary history and attenuation of myxoma virus on two continents. PLoS Pathog. 8, e1002950.

Kindler, E., Thiel, V., 2014. To sense or not to sense viral RNA-essentials of coronavirus innate immune evasion. Curr. Opin. Microbiol. 20C, 69–75.

Marra, M.A., et al., 2003. The genome sequence of the SARS-associated coronavirus. Science 300, 1399–1404.

Nakamura, Y., Gojobori, T., Ikemura, T., 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucl. Acids Res. 28, 292.

Pyrc, K., Berkhout, B., van der Hoek, L., 2007. The novel human coronaviruses NL63 and HKU1. J. Virol. 81, 3051–3057.

Pyrc, K., Jebbink, M.F., Berkhout, B., van der Hoek, L., 2004. Genome structure and transcriptional regulation of human coronavirus NL63. Virol. J. 1, 7.

Sharp, P.M., Li, W.H., 1987. The codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications. Nucl. Acids Res. 15, 1281–1295.

Shi, S.L., Jiang, Y.R., Liu, Y.Q., Xia, R.X., Qin, L., 2013. Selective pressure dominates the synonymous codon usage in parvoviridae. Virus Gen. 46, 10–19.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.

Vabret, N., Bailly-Bechet, M., Najburg, V., Muller-Trutwin, M., Verrier, B., Tangy, F., 2012. The biased nucleotide composition of HIV-1 triggers type I interferon response and correlates with subtype D increased pathogenicity. PLoS ONE 7, e33502.

van Boheemen, S., et al., 2012. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. mBio 3, http://dx.doi.org/10.1128/mBio. 00473-12, pii:mBio.00473-12.

van der Hoek, L., Pyrc, K., Jebbink, M.F., Vermeulen-Oost, W., Berkhout, R.J., Wolthers, K.C., Wertheim-van Dillen, P.M., Kaandorp, J., Spaargaren, J., Berkhout, B., 2004. Identification of a new human coronavirus. Nat. Med. 10, 368–373.

van der Kuyl, A.C., Berkhout, B., 2012. The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. Retrovirology 9, 92.

Van Hemert, F., van der Kuyl, A.C., Berkhout, B., 2014. On the nucleotide composition and structure of retroviral RNA genomes. Virus Res. 193, 16–23.

van Hemert, F.J., Berkhout, B., 1995. The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. J. Mol. Evol. 41, 132–140.

van Hemert, F.J., van der Kuyl, A.C., Berkhout, B., 2013. The A-nucleotide preference of HIV-1 in the context of its structured RNA genome. RNA Biol. 10, 211–215.

Wang, M., Zhang, J., Zhou, J.H., Chen, H.T., Ma, L.N., Ding, Y.Z., Liu, W.Q., Liu, Y.S., 2011. Analysis of codon usage in bovine viral diarrhea virus. Arch. Virol. 156, 153–160.

Watts, J.M., Dang, K.K., Gorelick, R.J., Leonard, C.W., Bess Jr., J.W., Swanstrom, R., Burch, C.L., Weeks, K.M., 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 460, 711–716.

Wong, E.H., Smith, D.K., Rabadan, R., Peiris, M., Poon, L.L., 2010. Codon usage bias and the evolution of influenza A viruses. Codon usage biases of influenza virus. BMC. Evol. Biol. 10, 253.

Woo, P.C., Huang, Y., Lau, S.K., Yuen, K.Y., 2010. Coronavirus genomics and bioinformatics analysis. Viruses 2, 1804–1820.

Woo, P.C., et al., 2005. Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. J. Virol. 79, 884–895.

Woo, P.C., Wong, B.H., Huang, Y., Lau, S.K., Yuen, K.Y., 2007. Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses. Virology 369, 431–442.

Wright, F., 1990. The 'effective number of codons' used in a gene. Gene 87, 23–29.

Zaki, A.M., van, B.S., Bestebroer, T.M., Osterhaus, A.D., Fouchier, R.A., 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N. Engl. J. Med. 367, 1814–1820.

Zhang, Z., Dai, W., Dai, D., 2013. Synonymous codon usage in TTSuV2: analysis and comparison with TTSuV1. PLOS ONE 8, e81469.

Zuker, M., Turner, D.H., 1999. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: Barciszewski, J., Clark, B.F.C. (Eds.), RNA Biochemistry and Biotechnology. Kluwer Academic Publishers, Dordrecht/Boston/London, pp. 11–43.